1. Event Structuring as a General Approach to Building Knowledge in Time-Based Collections

William Ribarsky

The University of North Carolina at Charlotte

Zachary Wartell

The University of North Carolina at Charlotte

Wenwen Dou

The University of North Carolina at Charlotte

Keywords

Time-based data collections, event structure, temporal structure, narrative, storytelling, GIS.

Abstract

Many kinds of data collections are time-based or can be collected in a temporal manner. There has been a desire in the geography and geospatial communities to put temporal behavior on the same footing as spatial structure and to develop a comprehensive geo-temporal information system. In many cases, temporal information refers to sequences of happenings in the world. To efficiently represent such temporal information, we present event structuring as a general approach to build knowledge in time-based collections. In this case an event is defined as a meaningful occurrence that has substantial impact on subsequent developments. A properly organized event sequence forms a narrative, or story. Such stories are powerful mechanisms for human understanding; not only are they in a form that make them easier to recall, but they lead to mental models that can be intuited, examined, and joined together into larger models. In this paper, the pro-

posed event structuring methods are not limited to geospatial data, but rather any types of time-based collections such as text corpora. We first provide the definition of event structuring, and then describe detailed examples of event structures built upon different kinds of data. Last, we raise the need for an event descriptive language in order to generate, organize, and compare event structures.

1.1. Introduction

There are many kinds of data collections that are time-based. These include many types of physical data, either from simulations or observations, text collections with temporal information embedded, and multimedia collections with time stamps, embedded temporal information, or references to events in times. In many cases, this temporal information refers to happenings in the world, either real, simulated, or imagined. An example of physical data is the results from simulating a hurricane, where a storm surge forms that, depending on the direction and wind fields of the hurricane, is followed by inundation of barrier islands and coastal areas, after which there is flooding, rain squalls, and damaging winds. This sequence of events forms a narrative, or story, of the hurricane. An example of a text collection is a set of histories and/or reports. Embedded in the texts are dates, names, places, and occurrences that can be organized as events in time. For a particular geographical region, there could be social, political, military, weather, and other events that when brought together in an overall organization would reveal relationships among the events and thus among the histories. An example of a multimedia collection is online news pages aggregated over time. When properly organized along the time dimension, the stories extracted from these pages can be aggregated into topic clusters, which typically show a burst of related stories after a motivating event (e.g., the tsunami in Japan) (Luo 2010). For a large event there is an ebb and flow of related stories and many sub-topic clusters.

Central to all these types of data is the idea of an event, which we define here as a meaningful occurrence in time. In this paper, we demonstrate that an event is a powerful organizing concept, giving meaning and structure to temporal information. When properly organized, a sequence of events with coupled topics could indicate a cause and effect relationship and a sequence with similar topics would indicate a trend. A hierarchical event structure could emerge with larger events encompassing distributions of smaller events. Additional meanings would emerge from the event sequencing and structuring itself. A properly organized sequence would tell a

story, as in the hurricane example above. These stories are powerful mechanisms for human understanding; not only are they in a form that make them easier to recall, but they lead to mental models that can be intuited, examined, and joined together into larger models.

To the extent that a data collection refers to happenings in the world, there is also usually a geographical reference. A GIS can be used to organize and make queries of the geographical references, but there is then the question of how to include the time structure. Time is a dimension, like the three spatial dimensions, and this fact can be used to provide a similar organization as in the 3D GIS. But time is also different, as will be discussed further below. Events, here with the expanded definition of meaningful occurrences in time and space, provide the additional organization that will lead to a full 4D GIS.

1.2. Defining Events, Creating Event Structures, Organizing the Time Dimension

To organize the time dimension in a way similar to the spatial dimensions, we employ the events defined above. Since the unstructured time dimension is unbounded, the events provide a scale. Time units can be centuries, decades, years, months, days, hours, seconds, and nanoseconds. Seasonal weather events have a scale of hours to days whereas climate events have a scale of decades to centuries. The scale is determined by the event category.

The idea of a motivating event provides further structuring. The motivating event for the Japanese tsunami and nuclear meltdown was the tsunami itself. The motivating event for the Israeli incursion into Lebanon in 2006 was the firing of rockets at Israeli border towns by Hezbollah. All sorts of sub-events, reactions, and responses flowed from this motivating event. In cases such as the latter case, there may be some dispute about the motivating event and when it occurred. However, topic-temporal analysis and/or spatial temporal analysis will find the strongest clusters of events and their associated motivating event(s). We have shown this to be the case for broadcast news story analysis (Luo 2010, Luo 2006). Since, for example, visual analysis tools that use this structure are exploratory, one does not expect to be given the single right description of a story but rather the main thread and associated relevant information from which one can make his own interpretation. There are other examples demonstrating that these

event structures can generally be generated automatically or semi-automatically. For example, we have shown that storm surge, hurricane windfield, and atmospheric events associated with air quality can be generated by identifying and tracking 2D and 3D features of interest in simulations or observational data (Yu 2010). Events associated with the development of new research themes and ideas can be identified in large collections of proposals or research papers using a combination of temporal and topic modeling (Dou 2011), though in this case information from outside the collections must be used to describe the events in detail (e.g., information about when a new NSF program was launched that led to the development of the new ideas).

The motivating events provide a hierarchical structure. A main motivating event can encompass subordinate motivating events. (In the case of the tsunami, these subordinate events would be the breakdown of cooling systems that led to successive core meltdowns, evacuations, emergency maintenance, and so on.) In addition, we use the fact that all these time structures have a beginning and an end. To this we apply a "narrative imperative", where we assume a story is told between the beginning and the end. As much as possible, we apply a shaping based on the event clustering (including spatial-temporal and topical modeling) and make an interpretation that brings out the story, since this will provide the most meaningful structure in time. We have found that this hierarchical structuring and narrative shape emerge in a variety of different types of temporal data; we posit that this is a general phenomenon. However, the best, most meaningful structuring cannot be done entirely automatically, though our experience is that much of it can be. Here the user's knowledge and reasoning can be inserted. One goal of a visual analytics interface would be to guide the user to do this at just the right point in the time structure. Because of the hierarchical structure, this can be done quickly at higher levels.

1.3. Events in Space: 4D GIS

Since Einstein, it has been realized that time stands on an equal footing with the spatial dimensions in the physical world. Although this equivalence is apparent at, say, the scale of the universe, time takes on a different character than the other dimensions at our earth-centered scale. Most significantly, time is unbounded at this scale whereas the other 3 dimensions are bounded. (The spatial dimensions are also unbounded at the scale of the expanding universe, but this is not the relevant scale for an earth-centered focus.) Thus, the spatial dimensions latitude, longitude, and alti-

tude are bound to the surface of the earth (and thin shells above and below the surface) since that is where almost all earth-focused detail lies. (The thin shell above the earth's surface extends, say, through the stratosphere, and that below the earth's surface to the deepest mines.) But unlike the spatial dimensions, time keeps unspooling, starting with the primordial earth and unfolding inexorably as the present becomes the past. One thing is for sure; each passing instant adds to the time dimension.

This temporal behavior has the effect of endless "stacking up" of occurrences over spatial regions. In fact, the GeoTime papers take advantage of this metaphor to provide a rich visualization of events over geography.

1.4. Events in a Narrative Structure

According to a report published by the International Data Corporation, information that was either created or replicated in digital form in 2007 alone was 281 exabytes, and the projected compound annual growth rate between 2008 and 2011 was almost 60%. Without a doubt, such information contains valuable knowledge regarding every aspect of our lives, such as history, social behavior, new scientific inventions, etc. However, given the overwhelming amount of information, it is nearly impossible to manually sanitize, extract meaningful events, organize, and analyze these digital collections. Although organizing such information based on content or meaning is important, creating event structures along time allows us to discover the historic evolution of the events, themes, and even ideas. One can construct narratives and stories that effectively summarize and make coherent large amount of information. Lawrence Stone (1979) defines narrative as: "the organization of material in a chronologically sequential order and the focusing of the content into a single coherent story, albeit with subplots. Narratives are analytic constructs that unify a number of past of contemporaneous actions and happenings, which might otherwise have been viewed as discrete or disparate, into a coherent relational whole that gives meaning to and explains each of its elements and is, at the same time, constituted by them (Griffin 1992, McCullagh 1978). Narrative permits a form of sequential causation that allows for twisting, varied, and heterogeneous time paths to a particular outcome" (Griffin 1993). Our process of event structuring is similar to narrative in that it's not just temporally aligned incidents; it is centered on events that signal the beginning of multiple thematically related incidents. And the structure of the events makes the connections and relationships between incidents easily inferable. Organizing unstructured information into an event structure allows

one to grasp the gist of massive amount of information. For example, an event structure constructed based on everyday news could clearly represent what is/was happening and how each event progresses throughout time. An event structure built upon social media such as tweets could provide an up-to-the-minute of what everyone is talking about around us and across the globe. However, with 140 million new tweets per day, constructing an efficient analysis has become a highly challenging problem.

1.1.1. Human-Computer Generated Linear Narrative

Ideally, the event-based narrative structures should be general enough so that their general aspects can be widely applied to different kinds of data sources, be it GIS or non-geospatial data. In addition, the narrative should be human-computer generated, since without automation it will not be scalable and without human input it will not be fully meaningful.

For a human-computer generated narrative, we have the following goals.

1. Create an interactive analysis program, *ING* ("InteractiveNarrativeGenerator") such that a human user, *P* ("Person"), using program *ING* can interactively compute a digital narrative, *N*, from an arbitrary dataset *D*.

2. Create a program, *NC*, ("NarrativeComparator"), that can autonomously compare and cluster large sets of these digital narratives generated via interactive computation using the *ING* tool.

Below we give our operational definition of narrative and show how our goals above have a perhaps not entirely superficial similarity to several fundamental theorems in computer science.

A digital narrative is a narrative encoded in digital media (whether it is encoded as data or a program is an independent issue). A human generated digital narrative is a narrative generated by a person, but without using a software tool that explicitly aims to semi-automate the creation of the final narrative structure. (So a narrative created by a person using a word processor, or electronic search and analysis tool such as Matlab, is still called a human generated narrative). In contrast, a human-computer generated narrative is a digital narrative generated by a software tool that explicitly extracts events in and/or helps the user interactively structure the digital narrative. Referring to our above, we desire to make an InteractiveNarrati-

veGenerator program that helps people generate human-computed generated narratives for arbitrary digital datasets.

At present, we limit ourselves to human-computer generated linear narratives rather than branching or recursive narratives. We briefly discuss some intuitive reasons here. Following Occam's Razor, one should generate the simplest narrative which explains the temporal dataset unless one can trade simplicity for greater explanatory power. We further take the following stance. When comparing two otherwise similar narratives of roughly the same length, a person will generally find linear narratives easier to understand than branching narratives or recursive narratives. This is consistent with the observation that introductory computer science students are typically introduced to imperative programming constructs starting with sequential statements, then branching, then loops and finally recursion. A major caveat is that narrative length is important as well. One can imagine two narratives (which happen to be proper algorithms) that have equivalent interpretations, but where the first one is linear and 100 sentences long and the second one is iterative using a loop construct and only 3 sentences long. The former would be easier to understand for a person with imperative programming knowledge. However, we contend that generating a short linear narrative whose sentences use a high-level of abstraction is more desirable than generating a more precise (i.e. using less abstraction) branching or recursive narrative.

1.5. Events in Non-Geographic Information Spaces

There is much time-dependent information that is not spatial or not strongly spatial (e.g., text or multimedia collections). The event structuring described above can be pulled out from the spatial dimensions and used alone for these types of data. It appears that the ideas developed for geographic time will apply as well to these time-dependent information spaces.

Take document collections as an example: similar to the 4D GIS, document collections contain bounded dimensions plus the unbounded time dimension. The words, for example, are bounded by a finite vocabulary (though it may slowly grow over time), and the organization of the words is bounded by a set of explicit rules – grammar. Eventually copying all documents, from the beginning of writing on paper, to digital space from a paper-centric should bring tremendous benefits. However, the growth of digital information is exponential, as described above.

Visual Representation and Analysis of Temporal Structures. Certain types of visualizations, such as ThemeRiver (Havre, 2000), are particularly appropriate for representing temporal information. ThemeRiver is initially designed to visualize thematic variations over time within a large collection of documents (Havre, 2000). ThemeRiver provides a macro-view of thematic changes in a corpus of documents over a serial dimension. It is designed to facilitate the identification of trends, patterns, and unexpected occurrence or non-occurrence of themes or topics. Figure 1 shows an instance of ThemeRiver constructed based on microblogs. Although the x-axis displays time in a linear manner, according to studies on perception (Kohler, 1947, attention is usually drawn to the sudden increase and decrease of "currents" within the river. Therefore, significant patterns such as "bursty themes" are easily discovered by users through exploring the ThemeRiver.

Let's take a look at a concrete example of how ThemeRiver can facilitate identifying the beginning of an epidemic spread. The data are provided by the VAST challenge 2011 committee. The goal of the VAST challenge is to push the forefront of visual analytics tools using benchmark data sets and establishing a forum to advance visual analytics evaluation methods (IEEE VAST Challenge, 2011. One of the tasks in the 2011 challenge is to characterize an epidemic spread in a metropolitan area. One of the datasets for this task is the microblog messages collected from users in that region. The question is when and where the outbreak started and whether it is contained. With more than 1 million microblog messages in the data set, it is impossible to manually sift through all the messages, not to mention that lots of noise (random microblogs) exists in the data. An event structure constructed based on the microblogs can provide both a summary of all messages and insights regarding specific events such as the epidemic spread that the city officials worry about.

In order to construct the temporal event structure, we first processed all microblogs and extracted 10 thematically meaningful topics using Latent Dirichlet Allocation (LDA). LDA is a generative model that represents the content of words and documents with probabilistic topics (Blei 2003). The LDA has several advantages comparing to the previous vector space models (VSM) widely used for text analysis, one of which is that each topic now is individually interpretable, providing a probability distribution over words that picks out a coherent cluster of correlated terms (Blei 2010). Among the 10 extracted topics, two of them (Figure 1) are highly relevant to illness, in particular flu-like symptoms. Having the topical summary of

all microblog messages, we further visualize the topical trend in ThemeRiver to reveal temporal patterns and events. In Figure 2, each "current" represents a topic with the color schema connecting the topical trend and the actual content of the topic. The x-axis is time, with each interval denoting 4 hours in a day. In this case, we are portraying the topical trends of the microblog data between April 29th and May 19th.

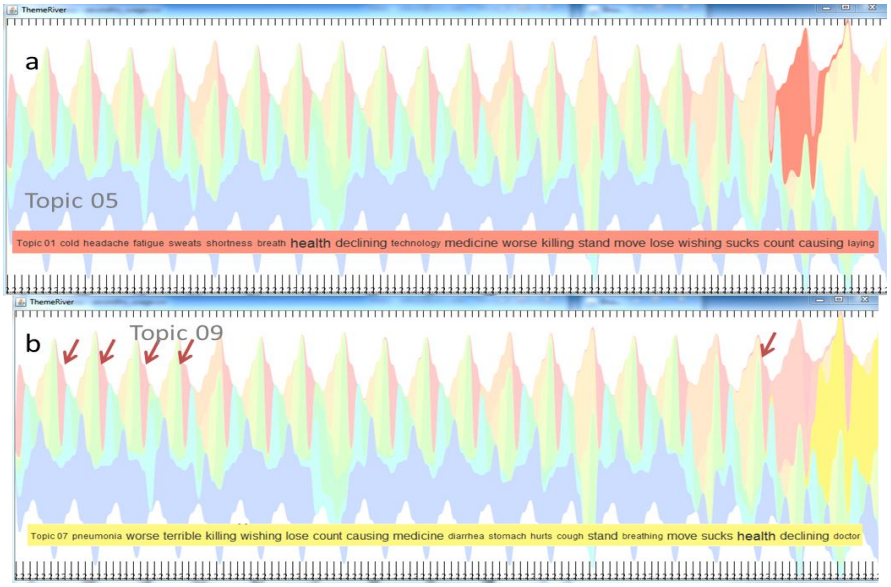

Figure 1: Two salient topical trends in the ThemeRiver regarding the epidemic spread. 1a: microblogs about flu-like symptoms such as "cold, sweats, headache". 1b: More severe symptoms such as "pneumonia, diarrhea" started appearing.

Given the ThemeRiver view, one can easily discover that there are repetitive patterns among several topics, such as users blogging about TV shows and commercials every night (topic 09 in red), and that lots of users like to talk about songs they love any time during the day (topic 05 in blue). However, what's really attracts attention is during the last 3 days the repetitive patterns suddenly broke. Instead the majority of the microblogs are about flu-like symptoms such as "cold, headache, fatigue, sweats, etc." (topic highlighted in Figure 1a) and "pneumonia, diarrhea, cough, etc." (topic highlighted in Figure 1b. These two topics signify exactly when the outbreak has begun. In addition, one can also infer a progression of the illness from cold and headache to more serious symptoms such as pneumonia, diarrhea and difficult breathing since all orange topic related microblogs appear before the yellow topic related microblogs.

The above example has shown that a ThemeRiver representation of temporal information can provide users with a quick summary of the microblogs. But more importantly, it directs users' attention to interesting patterns such as the sudden increase of microblogs on an epidemic spread. With such clear patterns, users could identify exactly when the epidemic started and how it has progressed. Such representation can be considered as a crude instance of the event structure. What would make the structure more complete is to use other sources of information such as news to label the ThemeRiver with motivating events, so that users could infer causal relationship between the news and people's reaction reflected in their microblogs. A powerful further advance would be to arrange the motivating events and important sub-events into a linear narrative, using the ideas described above. The whole arc of the epidemic could then be described in a coherent fashion. These considerations apply to all event-based temporal analyses.

Another data set provided by the 2011 VAST challenge committee is a text corpus containing new reports. If the ThemeRiver is properly labeled with relevant news information, the origin of the epidemic could be accurately discovered. We also applied the same topic modeling method to the news corpus and then filtered based on region and time to look for local news that might be related to the outbreak. Through our analysis, the most relevant incident we have found took place on May 17th, 2011, a news report of a truck accident on the interstate 610 bridge in the evening. The bridge leads over the main river in the metropolitan area and, as a result of the accident, the truck's cargo, probably containing some sort of chemicals, was spilled into the river. Tracing a few days back, on May 15th, a dangerous suspect who is member of the terrorist group "Paramurderers of Chaos" was arrested for trespassing near the loading docks at a food preparation plant in southwest part of the metropolitan area. Following this lead, we further discovered from the news data that the terrorist group had been planning a bioterrorism attack on the metropolitan area, which includes robbing equipment from a local university to manufacture dangerous microbes. As shown in Figure 2, with proper news events labeled on the microblogs data, one can infer causal relationship between an event and reactions to the event and begin to make an overall narrative.

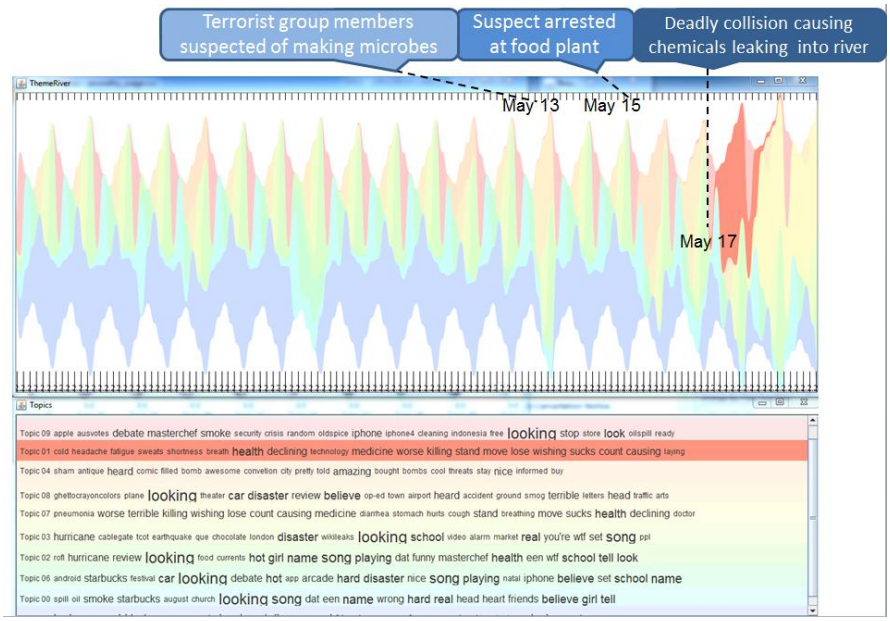

Figure 2: The microblog topical trends labeled with news information. On May 17th, right after a deadly collision involving a food truck that leaked chemicals into the major river in the city, people start showing flu-like symptoms. Following this lead, one can discover from the news that members of a terrorist group were manufacturing dangerous microbes and poisoned a local food plant. Through combing the microblogs with the news information, not only one can discover "what happened" with respect to the epidemic spread, but also what caused it.

Through combining different sources of information to augment each other, one can construct an effective event structure, which not only summarizes "what happened" but also allows inference of causal relationship between an event and subsequent outcomes. Above we used news reports to label information gathered from social media (how people react to certain news reports). Such structure could allow one to immediately identify what have caused the reaction and infer why people have reacted in a certain way. Similar ideas could be applied to the field of scientific research. For example, when visualizing the topical trends of scientific publications, other sources of information such as grant awards could be used to label the trends. Therefore one can infer the impact of the grant award on the evolution of scientific fields. The temporal scale of such analysis might be significantly different from analyzing a news corpus, which is more instantaneous, causing immediate splash in social media. The lag between the time one scientific proposal has been awarded and the time that similar

topics are seen in the form of publications could be more than a full year. Therefore, when constructing event structures, proper time scale should be carefully chosen. But, as we have noted above, this scale can be found from analysis of event patterns for a given category of events. In the case of scientific themes, for example, application of topic modeling to proposal and scientific paper collections would, upon analysis, reveal the spacing between an idea introduced in a proposal and subsequent development of that idea in papers.

1.6. Event Description Language for Linear Narrative

In order to generate, organize, and compare precise narratives, we need an event descriptive language. In this section we outline an event description language for linear narratives. We use object-oriented design to describe the language and temporal database terminology (Zaniolo 1997). We are continuing to investigate the literature on temporal logic (Øhrstrøm 1995), trace theory (Mazurkiewicz 1995), structural equation modeling (Pearl 2002), discrete-event modeling and simulation (Wainer 09) and narratology (Manford 2005). There is significant overlap between these domains and our goals.

We briefly review temporal database terminology. An instant is a single number, a 1-dimensional point in time. A period is a pair of instants. An interval is a single number, a 1-dimensional vector representing the displacement between two instants. Valid time is the historical time period during which a database fact is true. Transaction time is the period of time when a fact was entered into the database. Temporal databases and TAMs may support valid-time only, transaction-time only, or both. The latter is called a bi-temporal database. Decades of real-world usage indicate that bi-temporal databases should be provided because as temporal facts are gathered, changes, corrections, and filling in of omissions are inevitable and end-users inevitably want the ability to rollback the database to see the history of these change operations.

In our class hierarchy, an instant has a numeric value, a unit of measure, a calendar and a confidence descriptor. The latter three may be stored in an instant's tuple or computed. The calendar is a 1D temporal coordinate system. The confidence descriptor may indicate a confidence interval, a probability density function, or special value indicating either no error or that confidence information is not available. An instant's numeric value may be +infinite. For example, a database fact with an associated period

(0,+infinite) is interpreted as holding true from instant 0 through the rest of time.

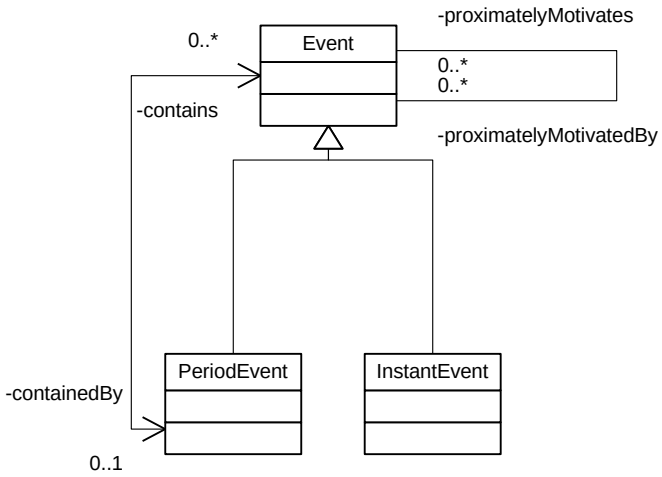

Figure 3: Abbreviated class hierarchy for our event structure

Next we define our event classes. A schematic of the class hierarchy is shown in Figure 3. An Event is an abstract class with 0 or 1 parent PeriodEvent objects (a forward declared class). PeriodEvent is an abstract class that has a valid time period and transaction time period. An InstantEvent is an abstract class that has a valid time instant and transaction time period. For example consider a sample from a digital thermometer such as (60°, 12:00PM 8/19/2011). This indicates a temperature of 60° was recorded at the given instant. A PeriodEvent also has a list of child Event objects. This parent-child structure defines a navigatable 1-to-N binary relation called TemporalContainment. The TemporalContainment relation induces a forest of tree structures on Event objects. There is a second bi-directional, navigatable N-to-N binary relation called the ProximatelyMotivates relation defined on the Event class. If Event A proximately motivates Event B, then A is an Event as a proximate cause of B. We restrict the proximatelyMotivates induced graph to be a directed-acyclic graph (DAG). Further, an Event A is said to "motivate" an event C if there is a path through the proximately motivates graph from A to C.

Various constraints must be maintained between the proximately motivates and temporal containment relations to avoid semantic inconsistencies. For example, a PeriodEvent's valid time period must contain the valid time periods of all child PeriodEvent's and the instant of all child InstantEvent's.

The structure of the temporal DAG can be created from an analysis of the motivating events and relevant sub-events, which can be derived from a variety of temporal feature analysis methods, as described above. The duration of events and their children (sub-events), appearance and disappearance of objects, cause-effect relations, and other temporal features can be described in the rich narrative descriptor language.

1.7. Towards a GTIS and TIS

For some time, there has been a desire in the geography and geospatial communities to put temporal behavior on the same footing as spatial structure and to develop a comprehensive geo-temporal information system. This need has become more acute because of the explosion in the number of compact, relatively inexpensive devices that make possible widespread, repeated measurement in the environment over time. (Repeated collection of airborne LIDAR data over wide areas is just one example.) That geo-temporal structuring is still an open problem is reflected in the recent report outlining important challenges from the National Academy of Sciences (NAS 2010).

Thinking in terms of a 4D GIS, as mentioned at the beginning of this paper, is one way of approaching this problem. However, as indicated above, time is different than the spatial dimensions, and therefore it is more accurate to speak about a *geo-temporal information system* (GTIS). Events can then be derived for each category of data in the system (e.g., geologic, climate, human history, weather, etc.) and the distribution of events over time provide the relevant scale and the periods for each category. Each of these event categories will have its own hierarchical structure and its own forest of trees, to which can be applied the human-computer narrative organization described above. GIS concepts can then be brought into the GTIS structure. For example, the idea of layers can be introduced. Each category could be a GTIS layer, which could be turned on or off as desired. A weather event layer could be overlaid on a human activity layer. One would then see the correlation, spatial and temporal, of these categories. There might be interesting cause-effect relations revealed between events in different categories. To encode these would require some interconnections of the originally independent categorical forests of trees. It would also be good to replicate in the temporal domain some of the GIS symbology and data organization. Of course, much of this grew out of hundreds of years of cartographic tradition. Road and political boundary vectors, for example, came from cartography and from the ongoing need to depict and

use this information efficiently. Not nearly as much work has been done for the temporal dimension nor is there a generally agreed upon set of symbols and data needs. The framework we have described here can be the basis for developing these things. For example, timelines annotated with key events are widely used to give an overview of history, military events, geologic events, key cultural moments, and so on. Having a structure within the GTIS to efficiently produce these would be a powerful thing.

To show the flexibility of such an approach, we recently described how the whole story of a city might be aggregated in a collection of hierarchical temporal event structures connected to a GIS structure (Ribarsky 2011). We chose Rome because of its 2500 year history as a major center of Western civilization. One can then consider architectural, military, political, cultural, ethnographic, weather/climate, disease spread, and other histories, each with its own event hierarchy embedded at certain levels in the overall GIS structure. For Rome there are massive amounts of documentation, including texts and images, for this collection of histories. Moreover, there is a substantial and growing digital archive. But all this detail has never been brought together into an integrated, whole story of Rome. It is clear that new causes, effects, and relationships would be discovered if this were done. We described in this paper how interactive timelines could be set up for these overlapping histories, so that major historical events could be unfolded into their sub-events and so that selection of any events on the timeline would reveal the details of the geographic extent (or how something like urban demographics, for example, developed over time). When considering two or more histories together, key points and patterns of correlation become evident. The event and narrative structuring described in this paper provide a rich, meaningful, and effective organization for comprehensive histories like this.

As discussed in previous sections of the paper, there are many information sources that are richly temporal without having a strong spatial component. Document collections which are augmented over time (such as proposal, research paper, and report collections) often fall into this category. For these collections, a *temporal information system* (TIS) would be appropriate and very useful. It seems to us that the event structuring and narrative approach we have described here can be brought to these types of information. Of course, the nature of the narratives and time scales of the events might be different. We have already started doing this in our studies of research paper and proposal collections (Dou 2011).

Although work has been done that shows that data-driven, semi-automatic event discovery and narrative structuring can be developed for a variety of data (Luo 2010, Lu0 2006, Yu 2010), much work remains to be done. In particular, we must remember that the ultimate goal is to make events and narratives meaningful so that people can gather insights, develop actionable knowledge, and create powerful hypotheses and models. More work on automated processing to extract more meaningful results is needed. But we must keep in mind that the human is the final agent that reasons and attaches meaning. Thus improved interactive visualization techniques, especially ones that insert human intelligence at exactly the right points, are absolutely necessary.

Acknowledgments

This research is supported by DHS Center of Excellence - Natural Disasters, Coastal Infrastructure and Emergency Management (DIEM), DOE DE-FG02-06ER25733, and NSF 0915528.

1.8. References

Blei, D., Carin, L., and Dunson, D. (2010) 'Probabilistic Topic Models,' Signal Processing Magazine, IEEE , vol.27, no.6, pp. 55-65.

Blei, D., Ng, A., and Jordan M. (2003) 'Latent Dirichlet allocation,' Journal of Machine Learning Research, vol.3, pp. 993–1022.

Dou, W., Chang, R., Wang, X., and Ribarsky, W. (2011) ParallelTopics: A Probabilistic Approach to Exploring Document Collections. Accepted for publication, IEEE VAST.

Griffin,L. (1993) 'Narrative, Event-Structure Analysis, and Causal Interpretation in Historical Sociology,' American Journal of Sociology, vol.98, pp. 1094-1133.

Havre, S., Hetzler, B., and Nowell, L. (2000) 'ThemeRiver: Visualizing Theme Changes over Time,' Proc. IEEE Symposium on Information Visualization, pp. 115-123.

IEEE VAST Challenge. [Online] Available at: http://hcil.cs.umd.edu/localphp/hcil/vast11 (Accessed: 1 September 2011).

Jahn, M. (2005) Narratology: A Guide to the Theory of Narrative. English Department, University of Cologne. Version 1.8. [Online] Available at: http://www.uni-koeln.de/~ame02/pppn.htm (Accessed: 1 September 2011).

Kohler, W. (1947) Gestalt Psychology. New York: Liveright.

Luo, D., Yang, J., Krstajic, M., Fan, J., Ribarsky, W., and Keim, D. (2010) EventRiver: Interactive visual exploration of constantly evolving text collections. IEEE Trans. On Visualization and Computer Graphics, doi.ieeecomputersociety.org/10.1109/TVCG.2010.225

Luo, H., Fan, J., Yang, J., and Ribarsky, W., and Sato, S. (2006) Exploring Large-Scale Video News via Interactive Visualization. IEEE VAST, pp. 75-82.

Mazurkiewicz, A. (1995) 'Introduction to Trace Theory', in Diekert, G. Rozenberg, (eds). The Book of Traces V. Singapore: World Scientific. pp. 3-67.

McCullagh, C. B. (1978) 'Colligation and Classification in History.' History and Theory. Vol.13. pp. 267-84.

NAS (National Academy of Sciences) Workshop. New Research Directions for the National Geospatial Intelligence Agency (May, 2010, Washington, DC).

Øhrstrøm, P. and Hasle, P. (1995) Temporal logic: from ancient ideas to artificial intelligence. Springer.

Pearl, J. (2002) 'Causality: Models, Reasoning, and Inference.' IIE Transactions, 34(6), pp. 583-589.

Ribarsky, W., Sauda, E., Balmer, J., and Wartell, Z. (2012) The Whole Story: Building the Computer History of a Place. Accepted for publication, Hawaii International Conference on Systems Science (HICSS).

Sudkamp, T., (1988) Languages and Machines: An Introduction to the Theory of Computer Science (3rd Edition).

Wainer, G. (2009) Discrete-Event Modeling and Simulation A Practitioner's Approach. CRC Press. pp. 3–33.

Yu, L., Lu, A., Ribarsky, W., and Chen., W. (2010) Digital Storytelling: Automatic Animation for Time-Varying Data Visualization. Computer Graphics Forum, Vol. 29 (7) pp. 2271-2280.

Zaniolo, C., Ceri S., Faloutsos C., Snodgrass R.T., Subrahmanian, V.S., Zicari, R. Advanced Database Systems (The Morgan Kaufmann Series in Data Management Systems), 1997.